图书情报工作
第63卷 第3期 2019年2月

面向社交媒体的细粒度ADR本体的半自动构建方法研究

■ 魏巍[1] 傅维刚[2]
[1] 中南财经政法大学大数据研究院 武汉 430074 [2] 武汉大学信息管理学院 武汉 430072

摘要：［目的/意义］提出一个药物不良反应本体的半自动构建方法，构建的细粒度药物不良反应本体为利用社交媒体挖掘潜在的药物不良反应信号提供语义资源库。［方法/过程］首先，采用业务层次和语言层次相分离的设计理念，将用户在社交媒体中评论的药物不良反应表示成"对象要素－属性要素－描述概念"的形式。细粒度体现在社交媒体用户对药物同一不良反应描述概念表达的多样性上。然后，基于深度学习的思想，利用基于word2vec的描述概念候选词抽取算法自动地抽取出更多的描述概念候选词构建本体。［结果/结论］以糖尿病药物的建模实例表明，提出的细粒度药物不良反应本体的半自动构建方案，提高了本体构建的智能化水平，构建的细粒度药物不良反应本体为利用社交媒体挖掘潜在的药物不良反应信号提供语义资源库。

关键词：本体构建 ADR 社交媒体
分类号：G251
DOI：10.13266/j.issn.0252－3116.2019.03.014

1 引言

本体在哲学层面原指现实世界中事物的具体存在，引申到信息科学中的语义层面，则可将本体理解为通过描述概念及概念间关系构建的专业领域的知识表示体系。利用领域本体表示和组织领域知识，不仅利于领域知识的共享，而且基于本体的查询和推理机制更能将领域知识运用于人工智能解决实际问题，具有更高的研究和实用价值。

本体构建的真正难点是人工梳理相关知识的工作量大、开发周期长，而且单凭技术人员的知识储备和对概念间关系的主观判断，不易做到准确、全面。这种依靠人工构建本体知识体系的方式效率低、公认性差，成为制约领域本体发展的一个瓶颈。国外在提高本体开发效率方面的研究一直处于领先，提出的本体学习(ontology learning)技术就是其中最具代表性的研究成果。其目标是利用机器学习和统计等技术自动或半自动地从自然语言文本语料中提取领域概念和这些概念之间的关系，并将其用本体语言编码形成易于检索的本体。然而，在目前的技术条件下，实现完全自动的获取和处理知识还不现实，整个本体学习过程还是一个需要人工参与的人机结合的半自动构建过程。

2 相关研究

随着Web2.0时代的来临，社交媒体以前所未有的数据增长态势，积累了大量用户数据资源。用户经常搜索一些健康相关主题的社交媒体并在上面分享自己的用药体验，这些社交平台的出现，为药物不良反应监测提供了新的途径。自然语言处理技术与机器学习方法的应用为从社交媒体挖掘潜在的药物不良反应提供了必要的手段[1]。M. Yang等构建了一个利用半监督学习的文本分类方法从社交媒体数据中自动识别药物不良反应(Adverse Drug Reaction, ADR)信息的预警模型，可以帮助药品监管部门和制药公司识别社交媒体上可疑的ADR消息[2]；B. W. Chee等使用一种机器学习方法，基于在线健康论坛中提取的信息，将药物分类为FDA的观察名单和非观察名单[3]。制药公司也对来自患者即时的和直接报告的药物不良反应非常感兴趣，因为这些对药品上市后监督期间的早期报告，能够使他们及早地发现问题并采取措施，而免于受到更严重的法律诉讼和利益损失[4]。

健康医疗研究应当充分利用这些丰富的信息资源，V. Hunsel等的调查揭示了荷兰患者报告药物不良反应的动机，表明患者愿意在社交媒体上分享他们使用药品的经验[5]。这些用户生成内容(User Generated

作者简介：魏巍(ORCID:0000－0003－3580－8360)，讲师，博士，E-mail:503175355@qq.com；傅维刚(ORCID:0000－0003－4682－696X)，博士研究生。
收稿日期：2018－06－12 修回日期：2018－08－08 本文起止页码：108－114 本文责任编辑：徐健

魏巍，傅维刚. 面向社交媒体的细粒度 ADR 本体的半自动构建方法研究[J]. 图书情报工作,2019,63(3):108－114.

Content,UGC)的迅速涌现,已成为持续监测公共卫生资源和不良疾病事件的重要资产[6]。分析社交媒体网站上患者的叙述内容,对于评估患者感知的药物不良反应风险[7]和挖掘药物与不良反应之间的关系也是非常重要的[8]。研究表明,ADR 的患者报告对于可靠的药物警戒可做出重要贡献[9]。

构建药物不良反应领域本体可以在药物警戒或个性化健康医疗服务中发挥积极的作用。M. C. Cai 等建立了一个全面的 ADR 本体数据库 ADReCS(Adverse Drug Reaction Classification System),提供了 ADR 术语的标准化和层次分类[10];Julien S. 等构建了一个对 MedDRA 术语进行形式化描述的语义资源库 OntoADR,改进了 MedDRA 术语的检索和编码,并可以进行按需定制分组[11];密歇根大学的 Y. Q. He 等创建了不良事件本体 OAE(Ontology of Adverse Events),对医疗干预后发生的各种不良事件进行逻辑定义和分类,为不良事件的逻辑表示和分析以及决定其临床结果的重要因素提供平台[12]。国内学者李梅等构建了我国心血管药物不良反应中英文本体知识库,系统地表示心血管药物不良反应,为心血管药物不良反应的分析与知识发现提供基础[13];并基于本体对国内抗感染药物不良反应报告进行分析,促进抗感染药物的合理使用[14]。

然而,已有的药物不良反应本体是基于医学词典和专业医学知识库开发的,这些本体中涉及的是医学概念和专业术语,随着利用社交媒体挖掘药物不良反应重要性的提升,利用已有本体进行知识表示和推理已不再适应社交媒体用户多样化表达的需要,不利于潜在药物不良反应的发现。此外,利用社交媒体挖掘药物不良反应的过程中涉及的大量描述不良反应的词汇是用药者使用的非专业用语,而且不同的用药者对同一种不良反应现象的描述可能是多种多样的。在基于词典或本体的药物不良反应事件抽取研究中,这些未在词典或本体中出现的不良反应描述常常被错过和忽视,从而减弱了该种不良反应实际发生的强度,降低了医师、药品生产厂家、患者等对发生该种不良反应可能性的评价。因此,笔者提出的面向社交媒体挖掘的细粒度药物不良反应本体(Fine-grained Adverse Drug Reaction Ontology,FGADRO)可以解决上述问题,细粒度即体现在社交媒体用户对不良反应描述概念的多样化表达上。

本体构建人工参与工作量大,且不易做到准确全面,如何快速、全面地梳理领域概念,如何正确地匹配对象及属性,能否通过机器学习的方法进行辅助梳理,提高本体构建的效率和智能化水平,将是笔者主要解决的问题。

3 细粒度药物不良反应本体描述概念生成方法

笔者设计的本体模型基于业务层次和语言层次相分离的思想。首先用本体表示多层次的药物和不良反应分类,药物不良反应本体由要素和概念两个层面构成:要素是领域层次,一般描述药物不良反应领域对象及其属性,这个层面是与专业知识相关的,需要医学领域专家参与界定;概念是语言层次,描述基本的语言概念,例如服用药物后患者的生理反应、心理效应、对药效的评价等,这些语言概念是与专业知识不相关的语言资源,可以由技术人员基于用药者评论,对常见的语言概念进行收集和整理。这样,在进行本体设计时就可以将领域层次和语言层次相分离,领域专家可以专注在领域要素的维护上,而不需要再去关注语言表达上的细节,而语言概念的处理可以交由技术人员完成[15]。

将用户在网络健康社区中评论的药物不良反应表示成"对象－属性－评价"的形式。笔者建立的药物不良反应本体,依据药品分类体系,比如,药品分类中有一个类叫"DRUGS USED IN DIABETES",它又可以和对象要素、属性要素进行关联,对象要素包括 INSULINS AND ANALOGUES、Biguanides、Sulfonylureas、Alpha glucosidase inhibitors、Thiazolidinediones、DDP-4 inhibitors 等,属性要素是具体的身体部位或器官,对象要素和属性要素又同时可以和第三个层次"描述概念"相互关联,笔者将"描述概念"界定为患者在社交媒体上评论的服用某种药物引起的不良反应的描述词汇,如 nausea、cough、weight gained 等。通过"对象要素－属性要素－描述概念"进行相互关联和组合就构成了对"DRUGS USED IN DIABETES"这个本体挖掘表达式的设置,从中得到想要的本体描述形式,如"Insulin caused respiratory tract infection in the respiratory system"。所以,只要知道用户的观点是在描述某种药物服用后的效果和感受,就能够把这个描述分类到该种药物的不良反应类别下。这些描述概念由于来自广大网民,他们在健康社区中对服用药物后的评价语言五花八门,描述词汇多是非专业用语,甚至包含口语化的表达。这些语言概念关于人的情绪或人们对事物的评价,它们与专业领域无关,因此可以交由技术人员协助领域专家进行收集和梳理。

基于深度学习思想，笔者提出了基于 word2vec 的药物不良反应本体描述概念的抽取方法，通过机器学习辅助梳理药物不良反应细粒度的描述概念，实现药物不良反应本体的半自动构建。领域本体描述概念的抽取过程，主要包括：种子概念的提取、细粒度描述概念候选词抽取和细粒度描述概念候选词筛选 3 个环节。如图 1 所示：

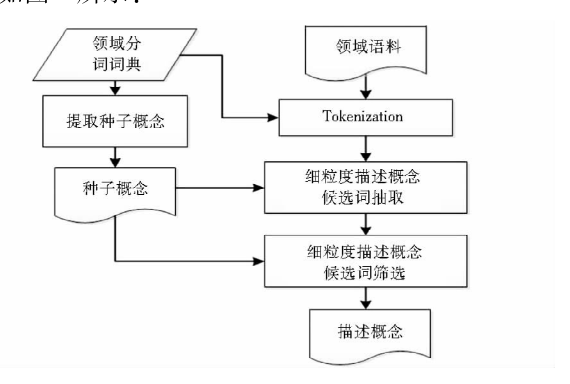


图 1　药物不良反应本体描述概念的抽取过程

3.1　种子概念的提取

本文中药物不良反应领域词典选择的是生物医学数据资源 SIDER（Side Effect Resource）。SIDER 是常用药物不良反应数据库，记录了多种上市药品及其不良反应信息。SIDER 中的不良反应名称已经映射到 UMLS（Unified Medical Language System）的超级叙词表中。领域知识通常围绕一些重要的概念展开，将领域中的核心概念称为种子概念[16]。例如在医学领域，“Cardiac disorders”“Endocrine disorders”“Gastrointestinal disorders”等术语都是由核心词“disorder”与其他词搭配而成。以种子概念作为中心词利用相应的算法，可以生成若干个扩展的领域概念。笔者利用药物不良反应领域词典，从相应药物已知的不良反应中挑选出最具代表性的核心词汇作为种子概念，以这些概念为基础，进行细粒度描述概念候选词的抽取和筛选。

3.2　基于 word2vec 的细粒度描述概念候选词的抽取

描述概念的获取是构建药物不良反应本体的关键环节，描述概念的自动抽取是指借助一定的技术手段，将反映某种药物不良反映特征或共性的词汇从一定规模的自由本文中抽取出来。本研究利用 word2vec 将描述药物不良反应的细粒度词汇映射为词向量，通过计算向量间的余弦值得到词汇间的相似度，搜寻与种子概念的相似度大于设定阈值的词作为药物不良反应描述概念候选词。

细粒度描述概念候选词的抽取过程如图 2 所示：首先使用 word2vec 工具对分词后的语料进行训练，得到词向量模型；然后以提取的药物不良反应种子概念作为输入词表进行初始化，利用该模型进行语义相关性计算，获得与输入词表的相似度大于设定阈值的词作为描述概念候选词。采取迭代算法扩充候选词集：将输入词表作为迭代变量，输出词表与输入词表的差集为输入变量，往复调用词向量模型扩充药物不良反应描述概念候选词集，直到符合迭代终止条件[17]。

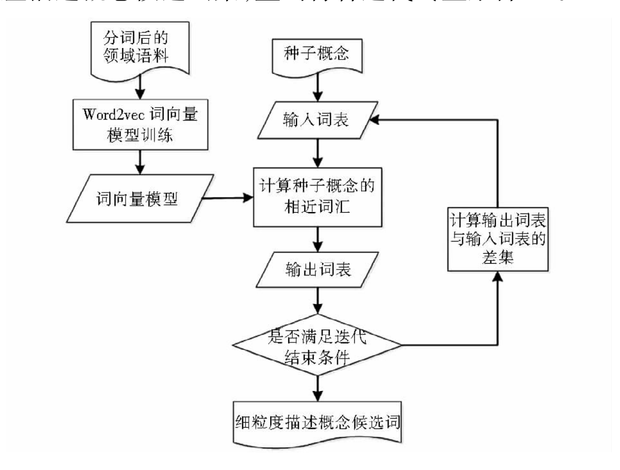


图 2　基于 word2vec 的细粒度描述概念候选词抽取流程

3.3　细粒度描述概念候选词的筛选

本体是一个专业领域的知识表示体系，本体的概念应是精炼的，规模也是确定的。因此，通过迭代算法抽取出的大量细粒度描述概念候选词应需要进行进一步的过滤，筛选出细粒度药物不良反应本体的描述概念。

本文的描述概念候选词筛选借鉴于娟等的思想，采用领域隶属度分析[18]的方法，该方法的基本思想是：如果某个候选词在前景语料（Foreground Corpora，Cf_k）中出现的概率比在背景语料（Background Corpora，Cb_k）中出现的概率高，且在前景语料中均匀分布，那么该词就有可能是该领域的描述概念。其中，前景语料是包含丰富种子概念的描述概念文档集，一般由若干标准化的领域文本组成；背景语料是与种子概念无关的其它领域的文档，用来与前景语料作对比，验证描述概念在其它领域中表现出的不同统计特性，一般由三个以上不同领域的若干文档组成。

描述概念候选词 t 与领域 D_k 的领域隶属度 DR 的计算公式如下：

$$DR_{t,k} = lg(TF_t) \times lg\left(\frac{P(t|Cf_k)}{P(t|Cb_k)}\right) \qquad 公式(1)$$

其中，$P(t \mid Cf_k)$、$P(t \mid Cb_k)$ 分别为候选词 t 在前景语料 Cf_k 和背景语料 Cb_k 中出现的概率。实际计算时，分别取估计值，具体计算公式如下：

$$E(P(t|Cf_k)) = \frac{TF_{t,k}}{df_k} \qquad 公式(2)$$

$$E(P(t|Cb_k)) = \frac{\sum_{Cf_i \in Cb_k} TF_{t,l}}{db_k} \qquad 公式(3)$$

$$TF_{t,i} = \sum_{C_j \in Cf_i} tf_{t,j} \qquad 公式(4)$$

其中,$TF_{t,k}$为候选词 t 在前景语料 Cf_k 中出现的频率, 为 Cf_k 的文档数目, 为 Cb_k 的文档数目,$tf_{t,j}$为 t 在文档 C_j 中出现的次数。

对每一个描述概念候选词分析其领域隶属度,将词以隶属度的降序排列,最后由领域专家依据本体构建的规模和该词的流行程度综合确认那些最能体现该不良反应特征的描述词汇作为细粒度描述概念。

4 细粒度药物不良反应本体的构建过程

在对常用本体开发方法进行比较的基础上,笔者采用基于知识工程[18]的开发方法对细粒度药物不良反应本体进行建模,半自动构建领域本体。领域专家参与,知识工程师人工设计本体架构;在提取药物不良反应的描述概念时,利用词汇的上下文语境生成词向量,通过机器学习的方法辅助梳理同义概念,提高本体构建的效率和智能化水平。细粒度药物不良反应本体的构建过程如下:

4.1 界定细粒度药物不良反应本体的范畴和目标

笔者开发的药物不良反应本体,是基于社交媒体中广大患者对药品使用后产生的不良反应的更细粒度的描述,涵盖更多患者评论的真实体验数据。细粒度药物不良反应本体是一个全面的药物不良反应本体知识库,不仅提供药物不良反应的标准化,而且提供药物不良反应各种描述概念的分级分类。它在生物医学和信息学研究中的应用远远超出简单的术语表。作为一个本体知识库,细粒度药物不良反应本体提供了一个直接计算药物不良反应相关术语间关系的机会,并提供了利用社交媒体挖掘潜在药物不良反应特征的线索。还通过寻求这些药物的共同特性,如药物的生理化学性质或蛋白质靶标结合,可揭示特定药物不良反应的分子机制,以协助未来更加合理的药物设计等。

4.2 列举细粒度药物不良反应本体中的重要术语和概念

在确定了领域本体范围的基础上,列举出药物不良反应领域涉及的相关重要术语和概念。将 MedDRA 和 UMLS 作为药物不良反应术语标准化的主要参考。这两个参考数据库在医学术语标准化方面做出了巨大贡献,他们的成果已得到业界的普遍认可。本文中药物不良反应本体的构建参照 UMLS(一体化医学语言系统)、MedDRA(国际医学用语词典)、SIDER(药物不良反应资源数据库),界定药物不良反应领域本体中的重要术语和相关概念,表 1 列举出了部分重要术语。

表 1 细粒度药物不良反应本体的重要术语

Disease	blood and lymphatic system disorders, blood and lymphatic system disorders, cardiac disorders, congenital, familial and genetic disorders, ear and labyrinth disorders, endocrine disorders, eye disorders, gastrointestinal disorders, hepatobiliary disorders, immune system disorders, infection and infestations, investigations, nervous system disorders, pregnancy, puerperium and perinatal conditions, psychiatric disorders, renal and urinary disorders, reproductive system and breast disorders, respiratory, thoracic and mediastina disorders, skin and subcutaneous tissue disorders, vascular disorders, surgical and medical procedures.
Drug (基于分类和作用系统)	alimentary tract and metabolism, blood and blood forming organs, cardiovascular system, dermatologicals, genito urinary system and sex hormones, systemic hormonal preparations, excl. sex hormones and insulins, antiinfectives for systemic use, antineoplastic and immunomodulating agents musculo-skeletal system, nervous system, antiparasitic products, insecticides and repellents, respiratory system, sensory organs, various...
Adverse Reactions	abscess, anaphylactic shock, arthralgia, back pain, bronchitis, tongue coated, cough, dysgeusia, dysmenorrhoea, dyspepsia, fracture, gingival bleeding, gingival hyperplasia, glossitis, tongue geographic, headache, hypersensitivity, hypertension, arthropathy, pain, hypoaesthesia, pharyngitis, paraesthesia, rhinitis, swelling…

4.3 定义细粒度药物不良反应本体类及类的层次体系

类用于描述抽象的实体对象,代表着一类具有共性的实例对象;类具有继承性并以层次结构的形式组织。层次描述了术语间的上位、从属关系以及纵向联系,更重要的是,层次结构允许在不同层次上进行计算,并支持将 ADR 逻辑链接到潜在的生理机能上。

依前文中本体表示模型所述,本研究中将本体挖掘表达式设置成"对象要素－属性要素－描述概念"的相互关联和组合,借鉴 MedDRA 和 WHO-ART 的层次结构,本研究中细粒度药物不良反应本体 FGADRO 的分层树包含 4 个层次:OE、PE、PT、Fg-PT。对象要素 OE 是依据药品分类体系对药品的分类;属性要素 PE 是系统器官水平的不利影响;细粒度体现在"描述概念"上,它包括两个层次:一个层次是 PT 层,表示特定的、唯一的和明确的 ADR 术语,它是必要的并且足以表征和区分其他 ADR 术语的单个医学概念。另一个层次是 Fg-ADR 层,描述医学概念细粒度的同义词、词

义变体、准同义词或子元素中 PT 层的扩展，笔者提出的面向社交媒体构建的细粒度药物不良反应本体即体现在此。利用前文第三部分提出的细粒度药物不良反应本体描述概念生成方法，通过机器学习辅助梳理同义概念，对反映同一药物不良反应的所有相关描述概念进行提取，提高药物不良反应本体知识库的全面性和科学性。

4.4　定义类的属性

这个阶段是本体构建的重要一步，将前一阶段建立的类以及类间的属性加以限制，表达了领域内更为完整的语义。对象属性和数据类型属性是 OWL 本体中的两种重要属性。对象属性用于描述两个类之间的相互关系；数据类型属性用于设定一个类自身独有的特性。通常情况下，定义类、类的层次结构和定义类的属性这两个步骤是一个不断重复的过程，先后顺序可以不加以区分，迭代进行。

4.5　生成实例

在定义了类及类的相互关系后就可以为本体创建实例了，一个类可以包含多个实例，一个实例也可能属于某一个类或多个类。依据前文第三部分细粒度药物不良反应本体描述概念生成方法，将生成的细粒度药物不良反应本体的描述概念以实例的形式表示和存储。创建药物不良反应概念类中的个体实体，也就是将表征和区分药物不良反应术语的单个医学概念在 FGADRO 类中以子类形式表示，而对同一医学概念细粒度描述的同义词、词义变体、准同义词等以添加实例的方式表示。

4.6　本体的检验和评价

本体构建的正确性是保证本体查询和推理有效进行的基础，因此，本体构建的最后一步是对所构建的领域本体进行完整性和一致性检验。本体推理机可用于识别本体中语法和语义冲突以及矛盾的知识。

5　建模实例

5.1　实验本体范围的界定

实验中将以糖尿病药物为例，构建细粒度的糖尿病药物不良反应本体。

实验数据来源于美国著名的药品评论公益网站 Ask a Patient，该网站为患者提供了一个交流和分享药物治疗经验的平台。选择糖尿病药物板块作为数据源，采集 39 种糖尿病药物的 17 682 条不良反应评论数据。将每种药物对应的不良反应评论数据分别存放于 excel 表中待作进一步处理。

5.2　基于 word2vec 的细粒度描述概念候选词的抽取

将数据集导入到 mysql 中，基于药物不良反应领域词典利用 Stanford NLP 对语料进行 Tokenization，包括词性还原、去停用词等处理，得到分词后的种子概念输入词表和药物不良反应语料词表。图 3 列出了药品 ACTOS 的种子概念输入词表。

1	2	3	4	5	6	7	8	9	10	11	12	13	14
Weight increased	oedema	Dyspnoea	Swelling	Somnolence	Fatigue	Lethargy	Arthralgia	Myalgia	Pain	Weight decreased*	Abnormal weight gain	Headache	Dizziness*

图 3　药品 ACTOS 的种子概念输入词表

采用 Python gensim 模块提供的 word2vec 工具包，将分词后的结果作为 word2vec 的输入对语料进行训练。训练结束后，得到语料中药物不良反应的描述概念候选词和其对应的词向量，通过余弦定理计算种子概念的语义近似词，进行排序并返回结果。通过 word2vec 提取的细粒度药物不良反应描述概念候选词的部分结果展示如图 4 所示：

weight increased描述概念候选词	余弦相似度
weight gain	0.913
fat	0.864
more weight	0.858
gain	0.701
pound	0.602

fatigue描述概念候选词	余弦相似度
tiredness	0.835
fatigue some	0.820
exhaustion	0.801
weary	0.797
exhausted	0.754

图 4　部分描述概念候选词的抽取结果

接下来，根据前文 3.3 节中公式对描述概念候选词进行领域隶属度的计算，将词以隶属度的降序排列提供给领域专家，进一步分析判断后确定 FGADRO 的描述概念。

5.3　细粒度药物不良反应本体建模

利用 Protégé 将糖尿病药物不良反应本体进行可视化。

药物不良反应本体类的确立基于医学领域知识，药学领域专家参与设计本体的类及类层次，该类层次结构主要包含 DRUGS USED IN DIABETES（糖尿病药物类）、Fine-grained ADR（细粒度药物不良反应类）、patient（患者类）3 个大类，每个大类下又包含相应的子类。糖尿病药物不良反应本体中涉及的重要属性包括“具有不良反应”“所用药物”等对象属性以及各种药

品及不良反应所对应的数据类型属性。图 5 显示了利用 Protégé 生成的细粒度药物不良反应本体部分的类及属性。

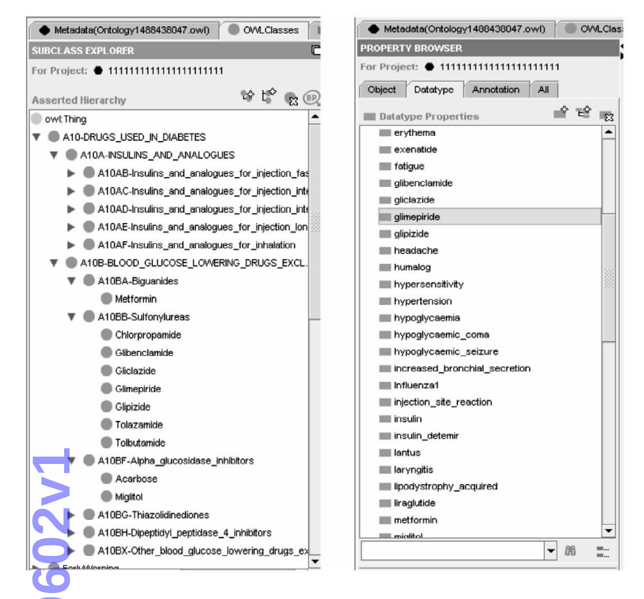


图 5　利用 Protégé 生成的糖尿病药物不良反应本体类和属性

将上文中筛选出的细粒度药物不良反应描述概念以实例的形式添加。也就是将表征和区分药物不良反应术语的单个医学概念在 FGADRO 类中以子类形式表示，而对同一医学概念细粒度描述的同义词、词义变体、准同义词等以添加实例的方式表示。

最后利用 Pellet 推理机对细粒度药物不良反应本体进行完整性和一致性检验。

本体的构建是一个复杂反复的过程，需要领域专家与知识工程师通力合作，对领域知识不断地完善和对细节不断的补充，并在不断的验证过程中，扩展完善本体。笔者构建的本体仅作为实验模型并用于验证后续研究中推理检测的有效性。

6　结论与展望

本文基于社交媒体数据源构建的细粒度药物不良反应本体采用业务层次和语言层次分离的设计理念，将用户在网络健康社区中评论的药物不良反应表示成“对象要素 - 属性要素 - 描述概念”的形式。细粒度体现在社交媒体用户对药物同一不良反应描述概念表达的差异和多样化上。本体构建的难点在于人工梳理相关知识的工作量大、开发周期长，而且单凭技术人员的知识储备和对概念间关系的主观判断，不易做到准确、全面。笔者提出的细粒度药物不良反应本体半自动构建方案，在细粒度描述概念的生成阶段，基于深度学习思想，利用机器学习等技术从社交媒体自然文本语料中提取描述药物不良反应的相关词汇，辅助梳理同义概念，并将其用本体语言编码形成易于检索的本体。一定程度上提高了本体构建的效率和智能化水平，构建的药物不良反应本体面向广大药物使用者，更能体现药物的真实体验数据，对利用社交媒体挖掘潜在的药物不良反应信号提供语义资源库。

此外，目前的研究只考虑了从特定社交媒体平台——网络健康社区收集的数据。网络健康社区相对其它社交媒体平台讨论的主题更为集中。使用其它流行的社交媒体平台(如微信和微博)数据，笔者提出方法的表现尚未得到证实。因此，对于未来的研究，可以探讨对于其他社交媒体(Twitter、Facebook 等)的适用性。

参考文献：

[1] 于跃. 基于大数据挖掘的药品不良反应知识整合与利用研究[D]. 长春:吉林大学,2016.

[2] YANG M, KIANG M, SHANG W. Filtering big data from social media-Building an early warning system for adverse drug reactions [J]. Journal of biomedical informatics, 2015,54(C):230 - 240.

[3] CHEE B W, BERLIN R, SCHATZ B. Predicting adverse drug events from personal health messages [J]. AMIA symposium, 2011 (4):217 - 226.

[4] VAN G K, DE G L, DE J, et al. Consumer adverse drug reaction reporting-a new step in pharmacovigilance? [J]. Drug safety, 2002, 26(4):211 - 217.

[5] VAN H F, VAN W C, PASSIER A, et al. Motives for reporting adverse drug reactions by patient-reporters in the Netherlands[J]. Europe journal clinical pharmacology, 2010, 66 (11): 1143 - 1150.

[6] YAN P, CHEN H, ZENG D. Syndromic surveillance systems[J]. Annual review of information science and technology, 2010, 42 (1):425 - 495.

[7] ABOU T M, ROSSARD C, CANTALOUBE L, et al. Analysis of patients' narratives posted on social media websites on benfluorex's (Mediator?) withdrawal in France[J]. Journal of clinical pharmacy & therapeutics,2014,39(1):53 - 54.

[8] LEAMAN R, WOJTULEWICZ L, SULLIVAN R, et al. Towards internet-age pharmacovigilance. Extracting adverse drug reactions from user posts to health-related social networks[C]// Proceedings of the 2010 workshop on biomedical natural language processing. Stroudsburg:Association for Computational Linguistics, 2010: 117 - 25.

[9] VAN H F, TALSMA A, VAN P E, et al. The proportion of patient reports of suspected ADRs to signal detection in the Netherlands: case-control study[J]. Pharmacoepidemiology drug & safety 2011, 20(3):286 - 291.

[10] CAI M C, XU Q, PAN Y J, et al. ADReCS: an ontology database for aiding standardization and hierarchical classification of adverse

drug reaction terms [J]. Nucleic acids research, 2015(43): D907 – D913.
[11] SOUVIGNET J, DECLERCK G, ASFARI H, et al. OntoADR a semantic resource describing adverse drug reactions to support searching, coding, and information retrieval[J]. Journal of biomedical informatics, 2016, 10(63):100 – 107.
[12] HE Y Q, SARNTIVIJAI S, LIN Y. OAE: The ontology of adverse events [J]. Journal of biomedical semantics, 2014, 5(1):1 – 13.
[13] 李梅. 我国心血管药物不良反应中英文本体构建与知识发现研究[D]. 长春:吉林大学,2017.
[14] 李梅,曹玉莹,张华吉,等. 基于本体的国内抗感染药物不良反应报告分析[J]. 药物流行病学杂志,2017 (2):115 – 119.
[15] 张佶. 传统银行的转型实战[EB/OL]. [2018 – 03 – 11]. https://www.geekbang.org/.
[16] 王红滨,刘大昕,王念滨. 基于遗传算法和种子概念的本体概念提取算法[J]. 系统工程与电子技术,2010,32(11):2465 – 2469.
[17] 闭炳华. 基于 word2vec 的数字图书馆本体构建技术研究[J]. 现代电子技术,2016,39(15):90 – 94.
[18] 于娟,党延忠. 领域特征词的提取方法研究[J]. 情报学报,2009,28(3):368 – 373.

作者贡献说明:
魏巍:设计研究思路与框架,撰写论文;
傅维刚:实验数据收集与分析,修改论文。

Semi-automatic Construction Method of Fine-grained ADR Ontology for Social Media

Wei Wei[1]　Fu Weigang[2]
[1] Big Data Institute, Zhongnan University of Economics and Law, Wuhan 430074
[2] School of Information Management, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] The semi-automatic construction method for the adverse drug reaction ontology is proposed. The constructed fine-grained ADR ontology provides a semantic resource library for exploiting potential ADR signals by using social media. [**Purpose/significance**] Firstly, based on the design concept that separates the business level and language level, this paper expressed the adverse drug reaction discussed in the network health community in the form of "object-attribute-description". The fine granularity is reflected in the diversity of describing the same adverse drug reaction. Then, based on the idea of deep learning, the word2vec-based description concept candidate word extraction algorithm is used to automatically extract more description concept to construct ontology. [**Result/conclusion**] The modeling example shows that the fine-grained ADR ontology construction scheme proposed in this paper can improve the efficiency and intelligence level of ontology construction. At the same time, the constructed fine-grained drug adverse reaction ontology provides a semantic resource library for exploiting potential ADR signals by using social media.

Keywords: ontology construction　ADR　social media